模糊规则归纳法及GDP主要影响因素分析

刘学生* 何 跃 贺昌政

(四川大学工商管理学院管理科学与工程系 成都 610064)

【摘要】使用一种新的自组织数据挖掘方法——模糊规则归纳法，根据GMDH技术，自动地从数据中提取模糊规则，形成更自然语言描述的模糊模型。阐述了模糊规则归纳法的算法原理及建模步骤，并给出了四川经济发展的建模研究实例，分析出影响GDP增长的主要因素，展示了模糊规则归纳法的突出特点。文中还对模糊规则归纳法和GMDH方法进行了比较，说明模糊规则归纳法进行因素分析的优越性。

关 键 词 模糊规则归纳法； 数据挖掘； 因素分析； GMDH方法； 国民生产总值

中图分类号 F201

一个模糊系统由一些具体说明输入输出映射关系的模糊规则所描述，模糊规则是写成IF–THEN形式，是一些专家知识的集合，预先将专家知识设计成规则的形式。无论是建立模糊控制系统还是模糊神经网络方法都要通过与专家对话预先确定一定数量的模糊规则[1, 2]，并在此基础上进行模糊推理。而自组织模糊规则归纳法用黑箱方法分析处理系统输入、输出变量之间的关系，以GMDH技术从系统已知输入输出数据自动地提取模糊规则[3]，这是该方法突出的特点。

1 模糊规则归纳法的原理

Zadeh发现，对真实世界的系统建模时用微分方程的建模途径有不足之处，故提出了模糊集合的思想。模糊模型可以用较自然的语言来描述复杂的系统。模糊规则归纳法(FRI)基于GMDH技术，自动地从数据中提取模糊规则，形成自然语言描述的模糊模型来描述复杂系统。FRI的执行过程就是应用黑箱的方法从数据中自动地建立模糊推理系统(指定输入输出映射关系的模糊规则的集合)，从而形成模糊模型的过程。执行n个输入变量的实值函数的模糊推理系统为[4]：

$$\boldsymbol{X}\rightarrow\text{模糊化}\rightarrow\text{模糊推理(规则归纳)}\rightarrow\text{逆模糊化}\rightarrow y$$

这里，$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$是多维输入变量，$y$是一个单变量的输出。

2 FRI的建模步骤

2.1 模糊化

模糊化是一个将输入变量$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$和输出$y$数值的观察值分别用隶属函数$x_{it}^{j}=m_{A_i^j}(x)$和$y_t^j=m_{B^j}(y)$转换为模糊向量$(x_{1t}^1,x_{1t}^2,\cdots,x_{1t}^m,x_{2t}^1,\cdots,x_{nt}^1,\cdots,x_{nt}^m)$和$(y_t^1,y_t^2,\cdots,y_t^m)$的过程。这里，$m$是每一个变量转换的语言变量的个数，本文使用$l$类型的隶属函数。

2.2 规则归纳

模糊规则写成IF–THEN的形式，通常是一些专家知识的集合，并通过专家的先知在规则基上确定。在黑箱方法中，目标是从系统的输入输出数据自动地产生这些规则，根据输出模糊集的个数m，产生m个静态或动态的模糊模型。

在运用GMDH的自组织FRI算法中，第一层的输入是由初始的输入变量的输入模糊集所描述。第一层的输入的个数是由模糊集的总数m和输入变量的个数n确定。即当产生一个静态的模型时，有

2001年9月20日收稿

* 男 24岁 硕士生

nm个输入神经元。对于一个动态的模型，有$n(L+1)m$个输入神经元，L是最大的时间延迟。

图1给出了一个多层的GMDH结构，其中每个神经元有2个输入(x_i^j, x_k^l)和一个输出(y^r)，如果$x_i^j \wedge x_k^l$，则$y^r(i,j,k,l)$。运用模糊逻辑，每个神经元由合取运算产生，即模糊集的AND联合。其最大最小推理为

$$y^r(i,j,k,l) = \min(x_i^j, x_k^l)$$

在第一层，模糊模型输出$y_t^r(i,j,k,l)(r=1,2,\cdots,m)$是在所有输入对$(x_i^j, x_k^l)$ $(i=1,2,\cdots,n,\ k \geqslant i,\ j=1,2,\cdots,m, l=1,2,\cdots,m)$的所有观察值$t$=1,2,…,$N$上进行估计。所有可能的组合产生以后，$F$个最好的2输入规则将被选出，用来作为第二层的输入，进一步产生有2，3或4个输入的模糊规则。模型筛选的主要准则为

$$\begin{cases} Q_1(i,j,k,l) = \sum_{t=1}^{N} | y_t(i,j,k,l) - y_t^r | \\ Q_2(i,j,k,l) = \sum_{t=1}^{N} (y_t(i,j,k,l) - y_t^r)^2 \end{cases}$$

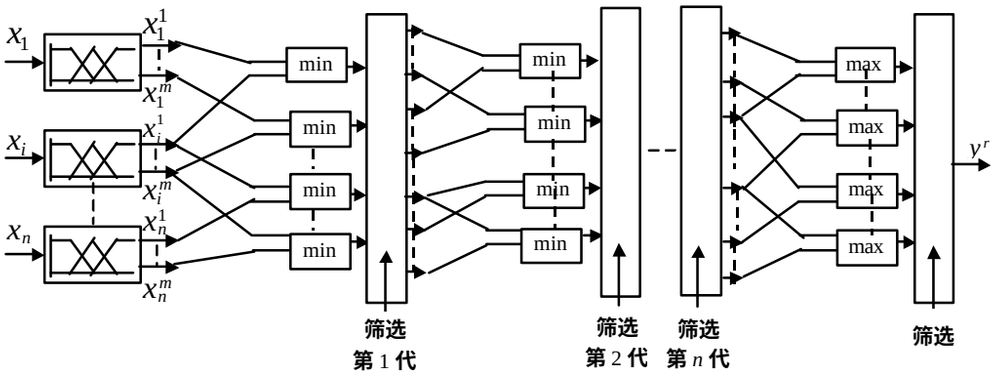

图1　规则的产生结构

随着层数的增加，当选择准则$Q_1(Q_2)$的值不减反增时，程序就停止(假设此时为第n层)。由一个转换函数$y_m^r(i,j) = \max(y_n^i, y_n^j)$产生所有第$n$层输出的$F$个最好模型的析取对。这里，$y_n^i, y_n^j$是第$n$层的输出，$i$=1,2,…,$F$，$j>i$。从准则$Q_1(Q_2)$的意义上看，最好的模糊规则将产生最终的模糊模型$y_m^r$。对$r$=1,2,…,$m$重复上面的过程，产生了最终的$m$个模糊模型$y_m^r$ $(r=1,2,\cdots,m)$。

2.3　逆模糊化

如果要求输出变量在原始数据库是可用的，则对它进行逆模糊化。用GMDH方法将估计的模糊输出$y^r(r=1,2,\cdots,m)$矢量转换为清晰值的原始区间。其转换公式为：$y = f(y^1, y^2, \cdots, y^r)$。如使用GMDH方法，由模糊输出$y^r(r=1,2,\cdots,m)$的线性组合来表示输出

$$y = a_0 + \sum_{i=1}^{m} a_i y^i$$

3　应用实例

本文以四川GDP增长的主要影响因素分析为例，以自组织算法中的FRI方法为主，同时运用了GMDH方法，并实施了两种方法的组合预测，将各方面因素都纳入系统，由计算机从数据的角度筛选变量并确认各变量的实际作用大小，详细演示了FRI的建模步骤。

文中用到了1985~2000年四川国民经济系统的27个变量的16组观察数据值[7]，其中x_1为国内生产总值(GDP)(亿元)；x_2为人口数(万人)；x_3为职工平均工资(元)；x_4为失业人数(万人)；x_5为从业人数(万

人)；x_6为城乡居民储蓄存款余额(亿元)；x_7为居民平均消费水平(元/人)；x_8为工业总产值(亿元)；x_9为固定资产总投资(亿元)；x_{10}为进出口总额(亿美元)；x_{11}为邮电业务总量(万元)；x_{12}为工农业总产值(亿元)；x_{13}为能源生产总量(万吨标准煤)；x_{14}为社会消费品零售总额(亿元)；x_{15}为普通高等学校在校生人数(人)；x_{16}为建筑业总产值(亿元)；x_{17}为货物周转量(亿吨·公里)；x_{18}为客运周转量(亿人·公里)；x_{19}为农林牧渔业总产值(亿元)；x_{20}为基本建设财政支出(亿元)；x_{21}为地方财政一般预算收入(亿元)；x_{22}为地方财政一般预算支出(亿元)；x_{23}为职工总工资(亿元)；x_{24}为金融机构各项存款年末余额(亿元)；x_{25}为金融机构各项贷款年末余额(亿元)；x_{26}为出口总额(亿美元)；x_{27}为进口总额(亿美元)。将16组数据分成学习集N_A＝13和校验集N_B＝3。集合N_A用来创建模型，集合N_B用来预测检验。

根据最优复杂度模型存在理论，最优复杂度模型很容易达到，可采用FRI和GMDH方法。

3.1 FRI方法

首先对数据进行模糊化。对原始数据取差分(使用 $\Delta x_{i,t}=x_{i,t+1}-x_{i,t}$，$i$＝2,3,…,27和 $\Delta y_t=y_{t+1}-y_t$)，用平均的Lambda型隶属函数，将输入向量 $\boldsymbol{X}=(x_2,x_3,\cdots,x_{27})$和输出$y$=$x_1$等距模糊化成3个语言变量组成的模糊向量$(\Delta x_2^1,\Delta x_2^2,\Delta x_2^3,\cdots,\Delta x_{27}^1,\Delta x_{27}^2,\Delta x_{27}^3)$和$(\Delta y^1,\Delta y^2,\Delta y^3)$。这样就产生了81个语言变量的数据集合。再利用这81个语言变量的集合创建一个具有4年滞后的模糊规则动态系统，即从324个语言变量的信息矩阵出发，在学习集N_A上分别对 Δy^i (i=1,2,3)创建一条模糊规则。每一条模糊规则将由几个先前未知的语言变量通过AND/OR/NOT操作符连接起来组成。

在预测了未来两年的模糊规则系统之后，利用相应的逆模糊化模型得到了最初的输出Y差分值的预测。逆模糊化的模型为

$$\Delta y_t=f(\Delta y^1,\Delta y^2,\Delta y^3)$$

四川GDP的规则如下：

R^1 如果t-3期的职工平均工资不是零(与样本的均值偏离较远)，并且t-2期的工业总产值不是正的(与样本的最大值偏离较远)，t-1期的客运周转量是零，t-3期的工业总产值不是零，则GDP就是零(与样本的均值很接近)。

绝对误差总和为0.050 0；平均绝对百分比误差为0.61%；近似方差为0.001 3。

R^2 如果t-2期的工业总产值是正的(与样本的最大值很接近)，GDP也是正的；如果t-1期的从业人数不是正的，并且t-1期的建筑业总产值不是负的，t-4期的居民消费水平不是正的，t-1期的社会消费品零售总额不是负的，则GDP就是正的(与样本的最大值很接近)。

绝对误差总和为0.010 0；平均绝对百分比误差为0.15%；近似方差为0.000 0所得到的逆模糊化的模型为

$$\Delta y=44.152\,843+179.707\,169\,\Delta y^1+452.924\,500\,\Delta y^2 \tag{1}$$

式中 Δy 为国内生产总值GDP的差分；Δy^1 为零_GDP；Δy^2 为正_GDP；平均绝对误差为3.21%；近似方差为0.003 0。由于 $\Delta y_t=y_{t+1}-y_t$，所以 $y_{t+1}=\Delta y_t+y_t$，由此得出对未来两年的预测值及其误差如表1所示。由于得到的FRI模型没有非滞后变量，所以非常适合于预测。

表1 FRI方法预测的结果及绝对百分比误差

时间/年	真实值	预测值	绝对误差/(%)
1998	3 580.26	3 560.666	0.547 4
1999	3 711.61	4 077.337	9.850 0
2000	4 010.25	3 993.938	0.407 0

3.2 GMDH方法

本文采用线性模型，利用N_A上的数据集合(13行，27列)和所选定的系统时间滞后4，由GMDH方法自动创建了如下信息

$$x_{1,t} = f(x_{2,t}, x_{3,t}, \cdots, x_{27,t}, \cdots, x_{2,t-1}, x_{3,t-1}, \cdots, x_{27,t-4})$$

用所得到的系统模型来预测未来2年的输出y，应用GMDH方法创建的具有最优复杂度的线性动态模型如下

$$y = x_1 = -310.239\,899 + 0.000\,167\,x_{11,t-1} + 0.049\,740x_{13} + 1.301\,928x_7 + 0.347\,754\,x_{25,t-2} \quad (2)$$

其中，绝对误差总和为0.010 5；平均绝对误差为0.11%；近似方差为0.000 0；预测结果如表2所示。

表2　GMDH方法预测的结果及绝对百分比误差

时间/年	真实值	预测值	绝对误差/(%)
1998	3 580.26	3 534.899 2	1.267 0
1999	3 711.61	3 983.107 0	7.314 8
2000	4 010.25	4 155.183 0	3.614 0

3.3　应用两种方法组合预测

应用GMDH方法创建的组合预测模型为

$$y = 1.009y_1 - 0.009\,544y_{2,t-1} - 2.016 \quad (3)$$

式中　y_1为GMDH模型；y_2为FRI模型；绝对误差总和为0.008 0；平均绝对误差为0.08%；近似方差为0.000 0。组合预测的结果如表3所示。

表3　组合预测的结果及绝对百分比误差

时间/年	真实值	预测值	绝对误差/(%)
1998	3580.26	3 532.237 6	1.341 3
1999	3711.61	3 982.231 9	7.291 2
2000	4010.25	4 150.894 0	3.507 1

3.4　模型分析

3.4.1　FRI方法

从逆模糊化的模型(1)可以看出，真实的输出与规则R^1和R^2有关。从以上的规则可以看出，影响四川GDP增长的因素有职工平均工资、工业总产值、客运周转量、从业人数、建筑业总产值、居民平均消费水平和社会消费品零售总额。由此可以得出如下结论：

1) 大力推进工业产业结构的优化升级，努力增强工业经济的整体实力和竞争力，对四川GDP的增长尤为重要。由上面的规则(尤其是规则R^2)可知，工业总产值在四川GDP增长中处于主导地位。四川省统计局最近公布的统计公报表明：四川省工业总产值对四川国内生产总值的贡献率接近50%，这与FRI得出的工业总产值在四川GDP增长中处于主导地位的结论相一致；

2) 加快发展建筑业，努力提高建筑业的总产值，有利于四川GDP的增长。四川省统计局最近公布的统计公报表明：建筑业在四川GDP增长中的作用也越来越显著。FRI模型能够把这个因素找出来，充分显示了FRI模型进行因素分析的优越性；

3) 努力改善人民生活，提高职工的平均工资，确保劳动者的充分就业，增加就业人数，对四川GDP的增长也有较大的推动作用；

4) 提高居民的平均消费水平，增强其消费信心，刺激消费需求，实现四川GDP的快速增长；

5) 加速发展旅游产业，交通运输业，带动客运周转量大幅度增加，有利于四川GDP的增长。

3.4.2　GMDH方法

由模型(2)可以看出，影响四川GDP的因素还有居民平均消费水平、邮电业务总量、能源生产总量和金融机构各项贷款年末余额。由此，得到如下的结论：

1) 提高居民的平均消费水平，能大大增加四川GDP。从模型(2)中居民的平均消费水平的系数1.301 928可以看出，四川的GDP增长主要受居民平均消费水平的影响；

2) 积极发展邮电事业，提高四川邮电业务总量，有利于四川GDP的增长；

3) 实现能源结构优化调整，努力增加四川能源生产总量，实现四川GDP的快速增长。从模型(2)中能源生产总量的系数是0.049 740，因而能源生产总量对四川GDP的影响来说比较大；

4) 积极增加信贷投入，在保障大中型企业信贷需求的同时，刺激消费信贷业务的蓬勃发展，能有效增加金融机构各项贷款年末余额，从而大大加速四川GDP的增长。

4 结 束 语

从上面的例子可以看出，FRI和GMDH方法都能为GDP影响因素的分析广泛地从数据中抽取一些信息。但是从预测的结果来看，FRI的预测效果要优于GMDH方法，并且，FRI对GDP影响因素分析的描述更接近于自然语言，具有更强的可解释性。当然，由于任何一个模型都是一种特定的抽象，只反映了现实的一些重要特征的某个或某几个方面，所以把FRI和GMDH方法所得的模型组合起来预测得到了较好的效果，也找到了影响GDP增长的较全面的因素，为经济发展战略的制定提供了有价值的信息。

参 考 文 献

1 Takagi T, Sugeno M. Fuzzy Identification of Systems and its Application to Modelling and Control.IEEE Transaction on Systems,Man,and Cybernetics,1985,15(1): 116-132

2 刘普寅, 吴孟达. 模糊理论及其应用. 长沙: 国防科技大学出版社, 1998

3 贺昌政, 梁元第, 王 绥. 数学建模导论. 成都: 成都科技大学出版社, 1997, 146-167

4 Mueller J-A, Lemke F. Self-Organising Data Mining[M]. Hamburg: Libri, 2000,131-137, 180-187

5 刘宝碇. 模糊准则决策过程及进一步研究. 电子科技大学学报, 1997, 26(增): 595-599

6 田益祥. 中长期预测模型的GMDH两水平算法. 电子科技大学学报, 1997, 26(增): 576-579

7 四川省统计局. 四川统计年鉴-1985-2000. 北京: 中国统计出版社, 1985, 2000

Fuzzy Rule Induction and the Analysis of Some Critical Factors Affecting the Increase of GDP

Liu Xuesheng He Yue He Changzheng

(Dept. of Management Science and Eng., School of Business Administration, Sichuan University Chengdu 610064)

Abstract In this paper we use a new self-organizing data mining method—Fuzzy Rule Induction(FRI), which extracts Fuzzy rules based on GMDH technique from the data autonomously to form the Fuzzy Model described in a more natural language. Some principles and processes of FRI method are developed in this paper. A model case about the economy in SiChuan Province of China is presented to analyzing some critical factors affecting the increase of GDP of SiChuan, while the related suggestions of policy are derived from this foundation. In addition, the comparison between the FRI and the GMDH has showed the advantage of FRI in anglicizing factor.

Key words FRI; data mining; the analysis of factor; GMDH ; GDP